Cognitive Media Theory (AFI Reader)

Audiovisual Correspondences in Sergei Eisenstein's *Alexander Nevsky*: A Case Study in Viewer Attention

Tim J. Smith

Cognitive film theory is an approach to analyzing film that bridges the traditionally segregated disciplines of film theory, philosophy and the psychological and neurosciences. Considerable work has already been presented from the perspective of film theory that utilizes existing empirical evidence of psychological phenomenon to inform our understanding of film viewers and the form of film itself.[1] But can empirical psychology also provide ways to directly test the insights generated by the theoretical study of film? In this chapter I will present a case study in which eye-tracking is used to validate Russian film director Sergei Eisenstein's intuitions about viewer attention during a sequence from *Alexander Nevsky* (1938).[2]

Knowing Where You Are Looking

One aspect of viewer cognition a filmmaker can influence in order to shape viewer experience is attention. Attention is the allocation of processing resources to a point in space, an object or visual feature resulting in an increase in information gathered by the senses.[3] Attention operates in all sensory modalities but the modalities most relevant to film are audio and visual. A filmmaker can direct visual attention to a specific element within a scene by cutting to a close-up; to an object within a shot by presenting it in sharp focus; or by directing audio attention to objects out of shot using off-screen sounds. By controlling what the viewer sees and what they do not see the filmmaker directs viewer comprehension and creates drama from a scene that might be rendered ambiguous if presented theatrically as a single unedited long-shot.

The composition of a shot — deciding what to include or exclude — is the clearest example of the way in which a director can manipulate viewer attention; but controlling where viewers look within a frame is also important, due to the limited visual acuity of the human eye. The non-uniform distribution of photosensitive receptors in the human retina means that we cannot see the whole visual scene in detail at the same time and must move our eyes to sequentially process the parts of the scene in which we are interested. Detailed colour processing happens in a small region (two degrees of visual angle) near the optical axis known as the *fovea* and image resolution drops rapidly as the distance from the fovea increases.[4]

Most encoding of visual information occurs when our eyes stabilize on a point in space (that is, fixate) and prioritizes the parts of the image projected near to the fovea.[5] To process a new part of the scene, the eyes must rotate so that the new target is projected onto the fovea; that is, to perform a *saccadic eye movement*. Visual sensitivity effectively shuts down during a saccade via a process known as saccadic suppression, in order to ensure that the rapid movement of light across the retina is not perceived as motion blur.[6]

Given that a cinema screen often subtends more than 35 degrees of visual angle,[7] viewer attention can only cover a small proportion of the screen at any one time. Therefore, if a filmmaker wishes to control what a viewer is perceiving at any moment during a film it is imperative that they are able to predict where a viewer will attend.

In "The Attentional Theory of Cinematic Continuity,"[8] I argue that careful manipulation of viewer attention within shots and across cuts is necessary to create the impression of a smooth flow of action across an edited sequence. I show how cinematographic and editing conventions, such as match-on-action editing (cutting between two views of an action just after the action begins), use natural attentional cues (for example, sudden onsets of motion) to attract and cue attention across cuts, creating minimal expectations about scene content that can be satisfied by shifts in viewpoint. Demonstrations of how viewer eye movements respond to natural attentional cues within film sequences and across continuity cuts support the attentional theory. However, such analyses are observational and based on hypotheses about how viewers may watch these sequences derived from post-hoc analysis of film sequences.[9] A stronger test of the attentional theory would be to demonstrate that a filmmaker's own intuitions about viewer attention may be validated through empirical investigation.

Audiovisual Correspondences

Filmmakers often abstractly discuss how they intended to manipulate viewer attention. Edward Dmytryck discusses the timing and movement of the viewer's eye following a character's exit from the screen,[10] while Walter Murch considers a viewer's "eyetrace" — "the location and movement of the audience's focus of interest within the frame"[11] — as one of the six criteria important for choosing the right cut.[12]

Unfortunately, descriptions of how filmmakers have manipulated viewer attention are often too abstract to facilitate empirical testing. The exception to this is Sergei Eisenstein's description of a sequence from *Alexander Nevsky*. In his book *The Film Sense*,[13] Eisenstein presents a detailed analysis of the famous "Battle on Ice" sequence from the film. In this sequence, the people's hero, Alexander Nevsky, leads a Russian army against the invading German knights across a frozen lake.

The scene ends with Nevsky's soldiers victorious and the German invaders swallowed by the freezing water as the ice cracks. The sequence of shots Eisenstein analyses precedes the conflict and represents the anticipation of the soldiers as they stare off into the distance at the enemy. Whilst mostly devoid of action, this sequence creates tension through what Eisenstein refers to as *audiovisual correspondences*; these "relate the music to the shots *through identical motion* that lies at the base of the musical as well as the pictorial movement."[14] Rising musical notes in the accompanying score by Sergei Prokofiev are said to correspond to rising "lines" in the pictorial composition. These pictorial lines can take various forms — including the arrangement of people, faces, objects or actions — but "the most striking and immediate impression [of correspondence with the audio] will be gained, of course, from *a congruence of the movement of the music with the movement of the visual contour* — with the graphic composition of the frame; for this contour or this outline, or this line is the most vivid 'emphasizer' of the very idea of the movement."[15]

To demonstrate the relationship between movement in the music and image Eisenstein presents a fold-out diagram at the end of the book (reproduced below as Figure 1). This diagram represents the temporal alignment of the movie frames to the musical phrases, the score, a diagram of the pictorial composition and — most critically for the topic of this chapter — a "*graph of the eye's movement* over the main lines...[of a shot] which 'correspond' to this music."[16] Eisenstein believes that to fully appreciate the film experience we must examine the combined impression the audio and visuals make on the mind of a viewer in the auditorium. The index of viewer cognition Eisenstein uses to gain this insight is eye movements. His graph of the eye's movement over the image is meant to represent the elements of the audio-visual continuity that exert the greatest pull on our attention. In effect, the graph is an idealized scanpath of the viewer's eye movements over the 2D surface of the image and over time.

The line of eye movements is presented continuously (except for a brief break in shot V) over the twelve shots, as if the shots were aligned side-by-side in a space in front of the viewer and their eyes entered a new shot in exactly the same position it left the previous shot. Of course, practically each shot replaces the previous shot on the screen via a cut, meaning that the horizontal progression of this line cannot be interpreted literally. However, Eisenstein's explanation of the diagram suggests that he intends the vertical movement of the graph to be interpreted as a direct representation of the vertical eye movements.

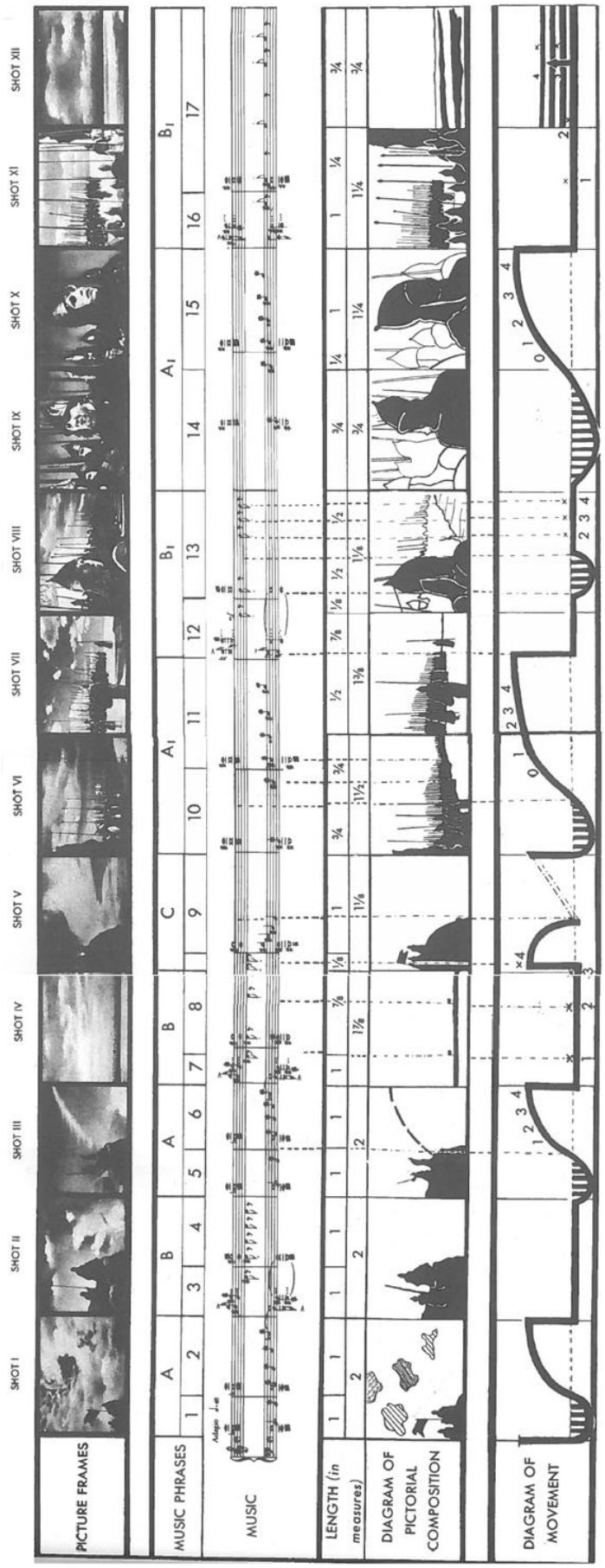

Figure 1: Diagram of audiovisual correspondences in Sergei Eisenstein's *Alexander Nevsky* (1938). Content of each row, from top: Picture frames, music phrases, musical score, duration (in measures), diagram of pictorial composition, and diagram of [eye] movement. Reprinted from *The Film Sense* (Eisenstein, 1943)

Eisenstein's belief in his ability to predict the collective viewing behavior of his audience is staggering. At the time he was making *Alexander Nevsky* the scientific investigation of eye movements was still very much in its infancy. In Chicago in the 1920s Guy Buswell developed an optical method for recording eye movements and used it to investigate reading and picture viewing, producing scanpaths of people viewing classic works of art such as Hokusai's *The Wave*.[17] Similar eye movement research did not occur in the USSR until after Eisenstein's death in 1948. During the 1950s, Alfred L. Yarbus, a biophysicist and professor in experimental psychology at the USSR Academy of Sciences in Moscow, pioneered several devices for recording eye movements and stabilising visual stimuli on the retina.[18] This research was published widely in Russian during the 1950s but did not gain international exposure until an English language translation of his monograph, *Eye Movements and Vision*, was published in 1967. Yarbus and Eisenstein were both very well connected within the Moscow scientific community, but it is unlikely that they ever met due to Eisenstein's untimely death before Yarbus began his eye movement research.[19]

The rich mixture of scientific, artistic, and political influences on Eisenstein's films and theories raise his work up above mere cinematic entertainments. His work may be understood as a synthesis of psychology, aesthetics and visual art,[20] and numerous pseudo-experiments can be found within his films. In *The Film Sense*, Eisenstein makes strong claims about how viewers should attend to the sequence from *Alexander Nevsky*, and these claims can be operationalized as testable hypotheses. To do this we first need to identify an appropriate measure of viewer attention and establish whether Eisenstein's claims are already supported by the existing psychological literature on film perception.

Attentional Synchrony vs. Idiosyncratic Gaze

One of the key assumptions of Eisenstein's analysis of the audiovisual correspondences in the "Battle on Ice" sequence is that all viewers will attend to the sequence in the same way: "The art of plastic composition consists in leading the spectator's attention through the exact path and with the exact sequence prescribed by the author of the composition."[21] The "spectator" here can be assumed to be all spectators. Is there evidence that viewers all attend to a movie in the same way?

In 1980, Barbara Anderson questioned the assumption of universal viewing behavior in an article entitled "Eye Movement and Cinematic Perception."[22] She referenced Alfred Yarbus's[23] demonstration of the influence of task on eye movements during static image viewing, and David Noton and Lawrence Stark's investigations into eye movement scanpaths.[24] These studies revealed the close ties between fixation location and perception in static scenes but also the large variation in

scanpaths between participants. Viewers may prioritise the same features of a scene, such as faces and points of high visual salience, but they do not necessarily attend to these features at the same time or in the same order. If the same behavior was observed during film viewing, such idiosyncrasies in attention would make it impossible for a film director to predict where their viewers would look. However, Anderson admitted that, "little experimentation [had] been done in this area with motion picture images" at the time she was writing.[25]

Since the 1980s, technological advances have made eye-tracking cheaper and more accessible but eye-tracking has rarely been used to investigate film viewing. Stelmach and colleagues were the first to record eye movements during film viewing.[26] When they collated the fixation points across all viewers they noticed something striking: the gaze of all viewers was tightly clustered within a very small portion of the screen area. This spontaneous clustering of gaze during film viewing has subsequently been replicated by numerous studies,[27] and has been named *attentional synchrony*.[28] The degree of attentional synchrony observed for a particular movie frame will vary depending on whether it is from a Hollywood feature film or from unedited real-world footage, the time since a cut and compositional details such as focus or lighting[29] but attentional synchrony will always be greater in moving images than static images.[30]

The key difference between a static image and a moving image is the inclusion of motion. In typical laboratory investigations of visual attention, sudden onset of motion has been identified as one of the most reliable ways to capture visual attention[31] In a computational analysis of the predictability of gaze by low-level visual features in movies — such as luminance, edges and motion — my colleagues and I found that by identifying points of high motion in a shot we could predict with reasonable reliability where viewers would fixate and when we would observe moments of attentional synchrony.[32] Points of high motion are said to be highly visually *salient* and involuntarily attract visual attention.[33] The reliability of viewing behavior during free-viewing of professionally produced moving images allows film directors to predict, with reasonable certainty, where their spectators will be fixating during most scenes.[34]

The "Battle on Ice" sequence chosen by Eisenstein to exemplify his use of audiovisual correspondences is problematic. We would predict that viewer gaze should exhibit attentional synchrony during this sequence, allowing Eisenstein to predict where viewers would fixate and whether their gaze followed the patterns transcribed in his diagram. However, Barbara Anderson's prediction that idiosyncrasies in gaze may render such predictions impossible may be valid, as Eisenstein chose this specific sequence because its mostly stationary nature allowed him to render it accurately on the printed page. In doing so, he may have omitted the motion critical for the creation of his intended correspondences between the film and viewer attention.

Cross-Modal Influences of Audio on Visual Attention

Eisenstein's predictions about how viewers will move their eyes during the "Battle on Ice" sequence in *Alexander Nevsky* differ to most previous investigations of viewer attention over static and dynamic scenes, in that he emphasizes the role played by audio in guiding visual attention. Most studies of attention to film treat hearing and vision in isolation,[35] but since the introduction of synchronized sound in the 1920s the film image has rarely been presented to the audience in isolation. Eisenstein's predictions may be invalid for silent film, but what about when accompanied by a soundtrack?

Investigations of visual and auditory attention traditionally occur in distinct and non-overlapping research fields. However, several key demonstrations of cross-modal effects exist. For example, the McGurk effect is a classic demonstration of how audio and visual perception is distorted by the presence of both sensory channels simultaneously.[36] The auditory presentation of the syllable /ba/, combined with the visual presentation of a face mouthing /ga/, will be perceived by an observer as mid-way between the two, /da/. The two channels combine to reconcile ambiguity in both channels, and create an integrated percept. Similar non-veridical audiovisual percepts are achieved by foley artists, who create sound effects — such as the sound of a body being stabbed or footsteps in snow — by using perceptually believable but mismatching objects to produce the sounds, such as stabbing a watermelon or walking on cornstarch.[37] Such audiovisual integration occurs automatically and very early in the perceptual processing of each sensory channel.[38]

In order to understand the influence audio may have on visual attention, we first need to establish the ways in which audio and vision can interact in film. "Sound may be diegetic (in the story space) or nondiegetic (outside of the story space). If it is diegetic, it may be on-screen or off-screen, and internal ("subjective") or external ("objective")."[39] If a sound object corresponds to a visual object on the screen, such as a barking dog (that is, a *diegetic on-screen external* source), the viewer will be able to shift their gaze to the source of the sound. If the barking was heard off-screen the viewer cannot overtly shift their attention to the sound source, but if the audio source is spatialized (through stereo or surround sound presentation) the viewer may covertly shift attention to the screen edge, cued by the audio to anticipate the appearance of the dog. Similarly, internal diegetic audio, such as a voiceover or access to a character's thoughts, may draw gaze to the associated character — if they are present on the screen — or covertly to the associated off-screen space.

These predictions of audio-cuing are supported by recent empirical evidence from various studies. The audio-cuing of off-screen space and its influence on viewer gaze has been demonstrated by Quigley and colleagues.[40] They recorded the eye movements of participants who were presented with static photographs of natural scenes, whilst a single ambiguous sound, such as

folding a piece of aluminum foil, was played in one of the four screen corners. Quigley and colleagues found that viewer gaze was biased towards the corner corresponding to the audio, even though there was no corresponding visual referent present in the image.

When the visual referent is present on the screen, such as the face of a speaker (that is, a *diegetic on-screen external sound* source), gaze will be biased towards the sound source,[41] and towards the lips if the audio is difficult to interpret.[42] Both the image and audio have independent influences on attention, but their combination has been shown to result in increased processing of both channels. The co-occurrence of a visual and sound onset will quicken detection of both relative to either presented in isolation and make the visual stimuli more perceivable.[43] Even spatially-uninformative sounds such as a simple "pip" can increase detection of a visual search target when it co-occurs with the onset of the target.[44]

Evidence for the increased activity of visual attention in the presence of audio has been demonstrated by eye-tracking experiments. Film clips presented with their original audio resulted in significantly greater attentional synchrony, larger saccade amplitudes, and longer fixation durations, than when the same clips were presented without audio.[45] This suggests that the audio was assisting all viewers in locating the most important area of the visual scene and maintaining fixation on it. Similar impact of the presence of audio on eye movements has been observed for close-up videos of people in conversation,[46] and a short film presented with four different versions of the soundtrack.[47]

But what about non-diegetic sounds? So far, I have reviewed empirical evidence indicating that sounds can direct our attention to, and increase our perception of, on-screen and off-screen visual objects. However, the audiovisual effect intended by Eisenstein in the "Battle on Ice" relies entirely on non-diegetic music to spatially cue visual attention. The literature on how music or auditory rhythm influences visual attention is very sparse. In a study by Vroomen and de Gelder,[48] participants listened to four tone sequences whilst they searched for a particular pattern of dots in a sequence of distractor patterns presented serially at fixation. The tones co-occurred with the onset of each pattern and were either four low-frequency tones, or three low-frequency tones with one high-frequency tone co-occurring with the target pattern. When the onset of the target was cued by the change in pitch, participants were significantly better at perceiving the target. Phenomenally, the sequence of patterns appears to momentarily pause, allowing greater detection. Vroomen and Gelder referred to the effect as the *freezing phenomenon*.

The rhythmical presentation of a series of pure tones has also been shown to entrain visual attention and quicken visual orienting and perception. Miller, Carlson, and McAuley[49] presented participants with a series of simple tones. When one of these tones coincided with the onset of a visual target, saccadic latencies to the target were significantly quicker than if the target appeared

just before or after the tone (+/-21ms). The rhythmical sequence also increased participants' ability to discriminate a target at fixation. These results, in combination with those of Vroomen and de Gelder,[50] suggest that covert visual attention can be entrained to an auditory rhythm, increasing perception at fixation and quickening responses to peripheral events. However, there is no evidence in the literature to support Eisenstein's more specific predictions about the influence of soundtrack on viewer gaze. His "diagram of movement" makes specific predictions about the vertical movement of gaze corresponding to increases and decreases of musical pitch. To the best of my knowledge, no empirical evidence for such directional effects of music on gaze exists in the literature.

Testing Eisenstein's Predictions

The lack of empirical evidence of non-diegetic, non-spatialized audio-visual effects on the direction of visual attention means that the only way to test Eisenstein's hypothesis is to directly record viewer eye movements during the "Battle on Ice" sequence. This approach could then be followed by more controlled studies that strip away features of the film in order to see which are critical for creating the intended effect.

Einsenstein states that, "The strongest impression . . . does not come from the photographed shots alone, but is an *audio-visual impression* which is created by the combination of [the] . . . Shots together with the corresponding music."[51] Therefore, we can formulate three hypotheses about viewer gaze during this sequence: (1) that all viewers will attend to the sequence in the same way; (2) that viewer gaze will follow the path described in the "diagram of movement" (Figure 1); and (3) that the presence of the audio is essential for creating the gaze behavior. In order to test these hypotheses, we presented two separate sets of participants the sequence with or without accompanying audio, whilst their eye movements were recorded with an eye-tracker.[52]

In order to test the three hypotheses stated above, gaze behavior in the original audio condition was compared to the silent condition. An initial method for comparing global eye movement behavior across the two conditions is to examine standard eye movement measures such as fixation durations and saccade amplitudes. If the audio was guiding eye movements to parts of the image, we might expect to observe larger mean saccade amplitudes and longer fixation durations, as was previously found by Antoine Coutrot and colleagues.[53] The mean fixation duration was numerically smaller in the audio condition (mean=437.14ms, Standard deviation, sd=121.98ms) than the silent condition (mean=445.81ms, sd=172.96ms) but this difference was not significant ($t(24)=-.148$, $p=.884$, not sig.). Average saccade amplitudes also failed to differ between the audio conditions (audio: mean=4.92 degs, sd=1.13; silent: mean=4.68 degs, sd=1.255;

t(24)=.509, p=.615, not sig.). This suggests that there were no global differences in how the two sets of participants controlled their attention whilst watching the sequence, and possibly suggests the primacy of the image in guiding visual attention.

The similarity of gaze in the two audio conditions becomes clear when visualized as dynamic heat-maps. A video of the gaze for both conditions (left=Audio, right=Silent) may be viewed on-line.[54] What is immediately striking when watching the synchronized audio and silent dynamic heat-maps is their similarity. Following cuts, both gaze groups snap to the same parts of the image and linger on similar regions during each shot's presentation. For example, in shot I of the sequence (Figure 1), both sets of viewers initially fixate the dark cloud at the screen's center, then saccade down to the figures in the bottom left as the fade-in gradually reveals them. What is also evident from the dynamic heat-maps is the frequency of saccades (2.27 per second on average) and the rapid way in which viewers explore the image over the long duration of each shot (mean shot duration = 7107ms), increasing the variance between individual gaze position as the shot continues. This pattern is characteristic of gaze during movie viewing.[55]

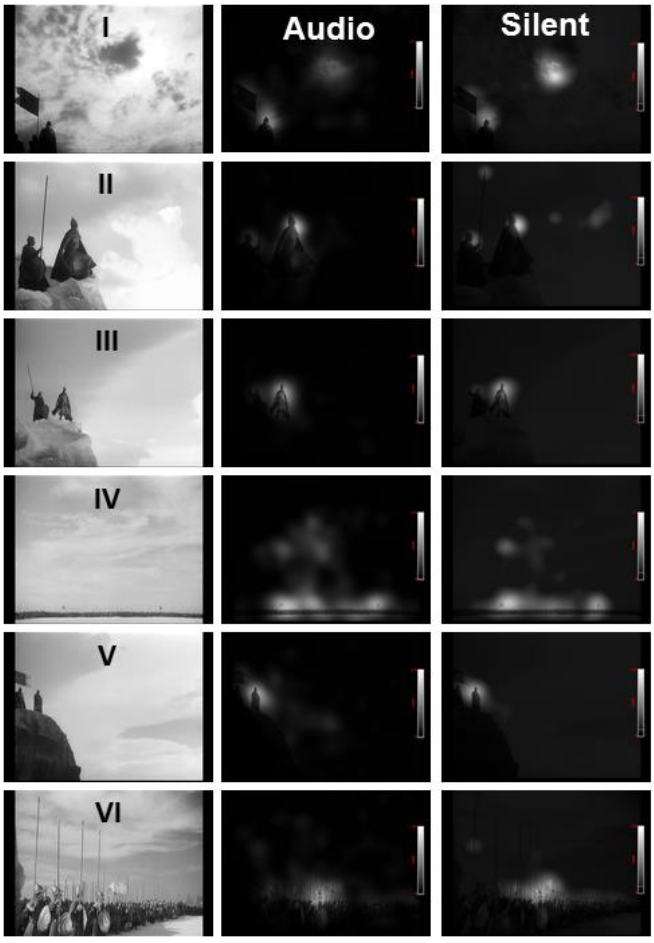

Figure 2: Left column=Shot sequence I to VI from Battle on Ice sequence of Alexander Nevsky (1938); Middle column = peak-through heatmap of fixations for the corresponding frame with original audio (Audio condition); Right column = peak-through heatmap of fixations for the corresponding frame without audio (Silent condition). Brightness of image indicates number of fixations.

In order to facilitate comparison of gaze in the audio and silent conditions, each shot from the sequence was converted into a peak-through heat-map (Figures 2 and 3). These static heat-maps represent the distribution of fixations during each shot's duration. Using these static heat-maps we can begin to look for qualitative differences between the audio conditions and look for signs of the movement patterns described by Eisenstein (Figure 1). Firstly, the distribution of fixations in the two audio conditions are very similar. The centers of interest (the brightest parts of the heat-map) match in both conditions and mostly correspond to the key features of pictorial composition identified by Eisenstein (Figure 1, row 3). For example, centers of interest in shot I are the central cloud, then the figures on the rock, in shot II they are the heads of the two figures in the bottom left; shot III are the same figures as in II; shot IV is the line of soldiers at the bottom of the screen, with particular interest in the two tall flags; and shot V returns interest to the two figures on the rock. Later shots (VIII, IX, and X) all demonstrate a gaze bias towards the faces of the characters shown in close-up. Our gaze initially snaps to their noses,[56] and then explores the eyes and mouth before shifting to other faces and figures. These centers of interest are very similar across both audio conditions.

Figure 3: : Left column=Shot sequence VII to XII from Battle on Ice sequence of Alexander Nevsky (1938); Middle column = peak-through heatmap of fixations for the corresponding frame with original audio (Audio condition); Right column = peak-through heatmap of fixations for the corresponding frame without audio (Silent condition). Brightness of image indicates number of fixations.

Static analysis of the distribution of fixations for each shot appears to mostly support Eisenstein's predictions about which features of the pictorial composition will capture viewer attention. However, Eisenstein's predictions are mostly concerned with the correspondence between the audio and the movement of attention between these centers of interest. This correspondence is most evident in the rising and falling pitch of the music, as demonstrated in this description of shot III and IV: "The first chord can be visualised as a 'starting platform'. The following five quarter-notes, proceeding in a scale upwards, would find a natural gesture-visualization in a *tensely rising* line . . . The next chord (at the beginning of measure 7), preceded by a sharply accented sixteenth, will in these circumstances create an impression of an abrupt fall."[57] Eisenstein claims that this rise and fall of the music matches the vertical movement of gaze across the image. To examine whether such a correspondence between music and vertical eye movements is observed in this sequence, the

average vertical gaze location (Y-coordinate) of all viewers over time was calculated and mapped against Eisenstein's diagram of movement (Figure 4).

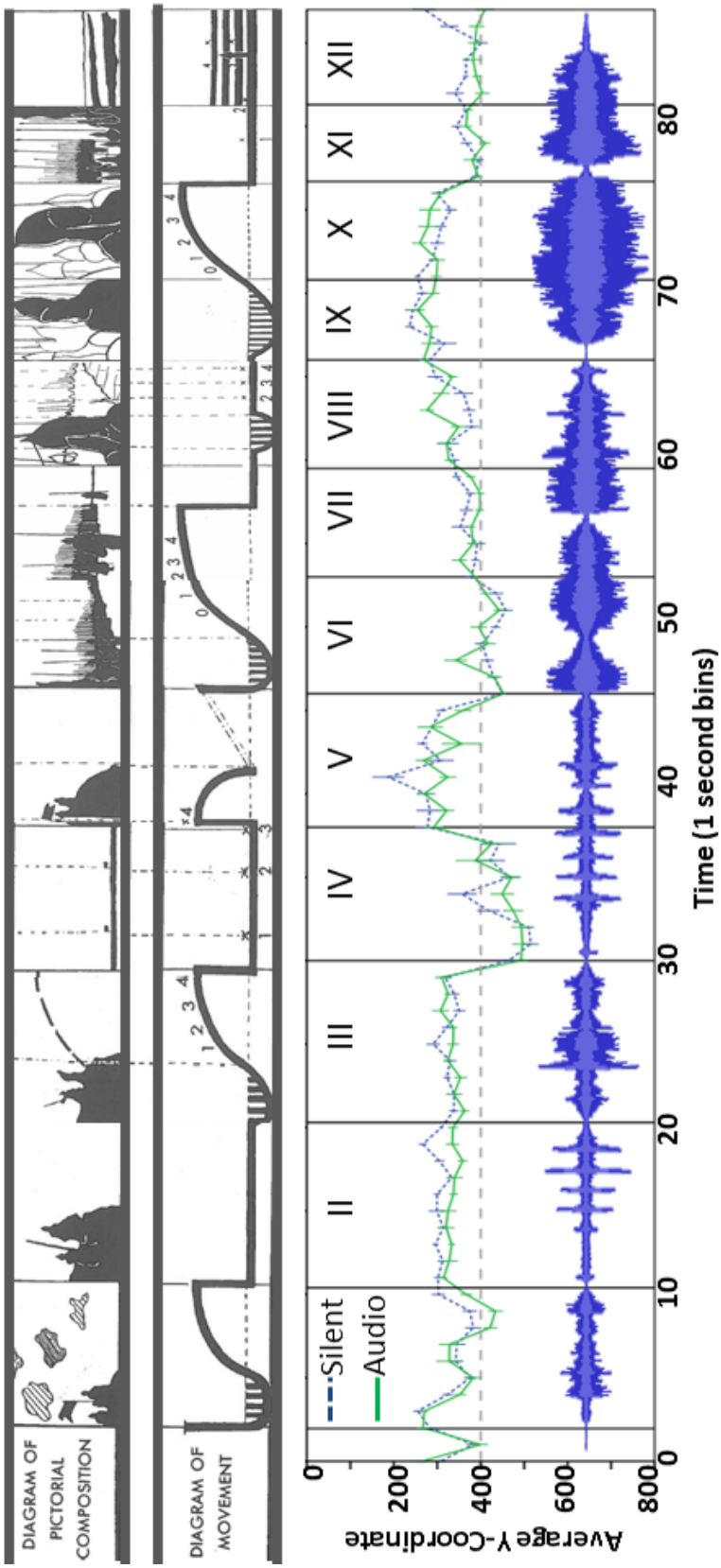

Figure 4: Top row of chart: Eisenstein's diagram of pictorial composition. Second row= Eisenstein's diagram of "movement" in each shot. Third row = Chart of average Y-coordinate of viewer gaze whilst viewing the sequence (0

coordinate = top of screen). Audio condition = Green solid line, Silent condition = Blue dotted line. Bottom row = audio waveform representing the soundtrack accompanying the sequence (amplitude of waves = loudness).

The vertical position of gaze in Figure 1 represents the mean across all participants within 1-second time-bins throughout the sequence. If viewer gaze was distributed evenly across the image, the average would remain at the screen mid-line (Y-coordinate = 400 pixels). Any deviation from this mid-line indicates a collective rise or fall of gaze and can be matched directly to Eisenstein's predictions via the aligned "diagram of movement."

An examination of the gaze for shots III and IV — which Eisenstein claimed was the most impressive audio-visual group — indicates a constant high-level of gaze in shot III, followed by a rapid drop to the screen bottom at the onset of shot IV. This drop directly mirrors the vertical movement predicted by Eisenstein in his "diagram of movement," and coincides with the musical drop in pitch from a B to a G sharp. The accompanying waveform shows how the eyes move during the silence between notes, and land in shot IV just as the isolated G sharp sounds (Figure 4; bottom row). The "tensely rising line" is absent from shot III prior to the cut, but this is probably due to the poor image transfer in the DVD copy used for this experiment. The arc of cloud Eisenstein expected to draw gaze up to top right of the screen is hard to discern in the digital version presented to participants (compare shot III in Figure 2 to shot III of the "picture frames" in Figure 1). However, even if this arc of cloud was visible in the image, such a static visual feature is unlikely to guide the gaze of all viewers at the same time.[58]

In a continuation of our examination of vertical gaze shifts, we see a sharp rise in gaze at the onset of shot V. This vertical movement coincides with the fifth repeated note from shot IV, and therefore does not create as clear an audiovisual correspondence as the fall from shot III to IV. Shot V is accompanied by a falling series of notes played on an oboe, followed by a silence and then a sharply accented rising note accompanied by piccolo. This accented note coincides with the first clear presentation of the massed troops and a sharp fall of gaze from the figures on the rock in shot V to the center of the troops in shot VI, creating a correspondence between gaze and audio.

The next sharp fall in Eisenstein's "diagram of movement" comes near the end of shot VII. Shots VI and VII depict the massed troops on the left of the screen, peering into the distance at the German troops. The musical sequence is identical to the sequence used in shots III and IV to denote a rise and then a sharp fall at the onset of IV. However, because the troops in shots VI and VII are aligned horizontally rather than vertically, Eisenstein claims that the accented sixteenth will create a "jolt" perspectively inwards, towards the horizon. Examination of the vertical gaze positions shows no change in gaze height, but watching these shots in the dynamic heat-map (see video) also shows no tendency to shift gaze towards the horizon coinciding with the sharp musical fall. We cannot rule

out the possibility that the musical sequence creates the perception of a movement in depth, but this is not evident in the gaze.

Shots VIII and IX are represented as mostly stationary by Eisenstein but culminating in a rise in shot X, followed by a sharp fall to shot XI as the images are accompanied by the same musical sequence as in shots III and IV. The corresponding vertical gaze position rises gradually from shot VIII to X, mirroring the rising notes in shot X, and then falls sharply at the onset of shot XI as the accompanying music also falls. This repetition of shot III and IV's audiovisual correspondences and gaze pattern ends the narrative unit depicting the soldiers waiting, and anticipates the dramatic and musical change about to occur after the mute images and audio of shot XII.

This analysis of the audiovisual correspondences and their manifestation in viewer gaze has confirmed several of Eisenstein's predicted rises and falls in gaze between shots (as in shots III-IV and X-XI), but also failed to support some of his more fanciful predictions about movements within shots (such as in shots VI-VIII). Due to the absence of image motion in this sequence, the only point at which Eisenstein could reliably predict where viewers would look was across cuts. Immediately following a cut, our attention is predominantly driven by low-level visual features,[59] by biases towards the screen center, and by social features such as faces and bodies.[60] By positioning such features at different heights before and after the cut, Eisenstein was able to create a reliable vertical shift in collective viewer attention. This movement may have been noted by Prokofiev when composing the score for the fully edited image sequence, and mirrored in the note transitions to create musical accents corresponding to the visual.

There is also an important question about the causal direction of the correspondences. Does the audio cause the corresponding gaze pattern or is the gaze caused by the pictorial composition and simply mirrored by the audio? Qualitative comparison of the dynamic heat-maps (Video), the static peak-through maps (Figures 2 and 3), and the vertical gaze coordinates (Figure 4), show no clear differences between the audio and silent conditions, suggesting the primacy of the image in guiding viewer gaze. Quantitative analysis of the vertical gaze coordinates in the audio and silent groups reveals a marginally significant difference in the overall vertical coordinates of gaze in the audio and silent conditions ($F(1,24)=3.291$, $p=.082$, marginal) with the silent condition often having higher gaze coordinates than the audio (see Figure 4). Both audio conditions show a significant effect of time ($F(87,2088)=14.204$, $p<.001$) but the difference between audio conditions does not change with time (that is, there is no interaction; $F(12,294)=1.263$, $p=.239$, not sig.). In general, the pattern of gaze is very similar across audio conditions and especially at the time points previously identified as moments of strong audiovisual correspondence (for example, transitions III-IV, V-VI and X-XI). Differences in gaze position across the audio conditions appear to mostly occur within

shots, but whether this is caused by the non-diegetic audio will have to be investigated in further studies.

Conclusion

This analysis of eye movement behavior during the sequence from Eisenstein's *Alexander Nevsky* has confirmed some of Eisenstein's predicted correspondences between movement of the musical score and movement of viewer gaze across the image. By formalizing Eisenstein's predictions made over 60 years ago as hypotheses testable using modern empirical methods (for example, eye-tracking), we have finally been able to confirm some of Eisenstein's intuitions and to provide empirical evidence that some filmmakers do indeed have the "uncanny facility to have your brain 'watch and note' your [own] eyes' automatic responses."[61] The greatest moments of attentional synchrony occurred immediately following cuts and were only preserved in shots with minimal pictorial details or clear faces or figures (for example, in shots I, II, III, V, VIII, IX, and X), whereas more complex shots (VI, VII, and XI) lead to less attentional synchrony. However, even in these complex shots the systematic gaze shifts across cuts were present. The lack of a significant change in gaze behavior when audio was removed suggests that in this instance the visuals had primacy in the creation of audiovisual correspondences. However, we cannot rule out the possibility that, by adding complimentary musical transitions, Prokofiev has vertically cued visual attention to covertly shift away from fixation, thereby influencing the perception of each shot. Such covert shifts are not measurable by eye-tracking and would require follow-on behavioral measures.

This study provides a template for future investigations into audiovisual correspondences in film, as well as related phenomenon in cross-modal attention and perception. For instance, the primacy of the image in the sequence examined here may not be present in other film sequences or in other styles of audiovisual material. For example, montage sequences are a common technique in Hollywood cinema and TV for portraying the passage of time or the completion of a task.[62] Montage sequences are structured by the accompanying music. The pacing of cuts and movement within each shot closely relates to the music, often at the expense of narrative or spatiotemporal continuity. As such, montage sequences may be expected to have primacy of the audio in guiding attention within a shot or guiding the choice and timing of a shot. The same audio primacy is clear in music videos. Here the choice of imagery and editing structure is generally subservient to the audio. Music videos are likely to demonstrate strong audiovisual correspondences and it would be interesting to see if these correspondences are reflected in viewer gaze as in *Alexander Nevsky.*

Finally, this study has also revealed the surprising dearth of research into cross-modal influences on attention using naturalistic dynamic scenes. A handful of recent eye-tracking studies have begun investigating the influence of diegetic sounds on overt visual attention,[63] but these fall far short of the level of insight into cross-modal effects claimed by sound designers and film theorists.[64] Future empirical investigations of the influence of diegetic and non-diegetic audio on visual attention and the cognitive and emotional responses to film would further our understanding of cinematic experiences, and support Eisenstein's claim that audio-visual correspondences are more than the sum of their parts.

---

1 Joseph D. Anderson, *The Reality of Illusion: An Ecological Approach to Cognitive Film Theory* (Carbondale: Southern. Illinois University Press, 1996); David Bordwell and Noël Carroll, eds., *Post-Theory: Reconstructing Film Studies*, Wisconsin Studies in Film (Madison: University of Wisconsin Press, 1996).

2 *Alexander Nevsky*, directed by Sergei Eisenstein and Dmitri Vasilyev (Moscow: Mosfilm, 1938).

3 Harold Pashler, ed., *Attention*, Studies in Cognition (Hove UK: Psychology Press, 1998).

4 John M. Findlay and Iain D. Gilchrist, *Active Vision: The Psychology of Looking and Seeing*, Oxford Psychology Series (Oxford: Oxford University Press, 2003).

5 John M. Henderson and Andrew Hollingworth, "High-Level Scene Perception," *Annual Review of Psychology* 50 (1999): 243-271.

6 Ethel Matin, "Saccadic Suppression: A Review and an Analysis," *Psychological Bulletin* 81, no. 12 (1974): 899-917.

7 THX. *THX tech pages*. 2012 [Viewing distance requirements]. Available from: http://www.cinemaequipmentsales.com/athx2.html (accessed February 13, 2012).

8 Tim J. Smith, "The Attentional Theory of Cinematic Continuity." *Projections: The Journal for Movies and the Mind* 6, no. 1 (2012): 1-27.

9 David Bordwell, *Observations on Film Art*, 2008. Available from http://www.davidbordwell.net/blog/2008/02/13/hands-and-faces-across-the-table/ (accessed February 4, 2013).

10 Edward Dmytryk, *On Filmmaking* (London: Focal Press, 1986), 27-33.

11 Walter Murch, *In The Blink Of An Eye: A Perspective on Film Editing*, 2nd ed. (Los Angeles: Silman-James Press, 2001), 18.

12 Bruce Block, *The Visual Story: Seeing the Structure of Film, TV, and New Media* (Woburn MA: Focal Press, 2001).

13 Sergei M. Eisenstein, *The Film Sense*, trans. and ed. Jay Leyda (London: Faber and Faber, 1943).

14 Eisenstein, *The Film Sense*, 136.

15 Ibid, 135.

16 Ibid, 138.

17 Guy Thomas Buswell, *How People Look at Pictures: A Study of the Psychology of Perception in Art* (Chicago: University of Chicago Press, 1935), xv; 198, including illustrations, plates and diagrams.

18 Benjamin W. Tatler, Nicholas J. Wade, Hoi Kwan, John M. Findlay, and Boris M. Velichkovsky, "Yarbus, Eye Movements, and Vision," *i-Perception* 1, no. 1 (2010): 7-27.

19 Eisenstein's theories of montage were heavily influenced by contemporary work on neuroscience and psychology and he even states in retrospect, "Had I been more familiar with Ivan Pavlov's teaching, I would have called the 'theory of the montage of attractions' the 'theory of artistic stimulants.'" In Sergei M. Eisenstein, *Notes of a Film Director*, trans. X. Danko (New York: Dover, 1970).

20 Oksana Bulgakowa, *Herausforderung Eisenstein* [*Challenge Eisenstein*] (Berlin: Akademie der Künste der DDR, 1988).

21 Eisenstein, *The Film Sense*, 148.

22 Barbara Anderson, "Eye Movement and Cinematic Perception," *Journal of the University Film Association* 32, nos. 1 and 2 (1980): 23-26.

23 Alfred L. Yarbus, *Eye Movements and Vision*, trans. Basil Haigh (New York: Plenum Press, 1967).

24 David Noton and Lawrence Stark, "Scanpaths in Eye Movements During Pattern Perception," *Science* 171, no. 3968 (1971): 308-311.

25 Barbara Anderson, 25.

26 Lew B. Stelmach, Wa James Tam, and Paul J. Hearty, "Static and Dynamic Spatial Resolution in Image Coding: An Investigation of Eye Movements," in *Human Vision, Visual Processing, and Digital Display II* (Proceedings of SPIE), ed. Bernice E. Rogowitz, Michael H. Brill, and Jan P. Allebach (San Jose, 1991), 147-152.

[27] Robert B. Goldstein, Russell L. Woods, and Eli Peli, "Where People Look When Watching Movies: Do All Viewers Look at the Same Place?" *Computers in Biology and Medicine* 37, no. 7 (2007): 957-64; Uri Hasson, Ohad Landesman, Barbara Knappmeyer, Ignacio Vallines, Nava Rubin, and David J. Heeger, "Neurocinematics: The Neuroscience of Film," *Projections: The Journal of Movies and Mind* 2, no. 1 (2008): 1-26; Tim J. Smith, "An Attentional Theory of Continuity Editing,", Unpublished thesis, (Edinburgh: University of Edinburgh, 2006), 400; Virgilio Tosi, Luciano Mecacci, and Elio Pasquali, "Scanning Eye Movements Made When Viewing Film: Preliminary Observations," *International Journal of Neuroscience* 92, nos. 1-2 (1997): 47-52; Ran Carmi and Laurent Itti, "Visual Causes Versus Correlates of Attention Selection in Dynamic Scenes," *Vision Research* 46, no. 26 (2006): 4333-4345.

[28] Tim Smith and John Henderson, "Attentional Synchrony in Static and Dynamic Scenes," *Journal of Vision* 8, no. 6 (2008): 773.

[29] For review, see Tim J. Smith, "Watching You Watch Movies: Using Eye Tracking to Inform Cognitive Film Theory," in *Psychocinematics: Exploring Cognition at the Movies*, ed. Arthur P. Shimamura (New York: Oxford University Press, 2013), 165-191.

[30] Tim J. Smith and Parag. K. Mital, "Attentional Synchrony and the Influence of Viewing Task on Gaze Behaviour in Static and Dynamic Scenes," *Journal of Vision* (2013): 13(8): 16.

[31] Jeremy M. Wolfe and Todd S. Horowitz, "What Attributes Guide the Deployment of Visual Attention and How Do They Do It?" *Nature Reviews Neuroscience* 5 (2004): 1-7.

[32] Parag K. Mital, Tim J. Smith, Robin L. Hill, and John M. Henderson, "Clustering of Gaze During Dynamic Scene Viewing is Predicted by Motion," *Cognitive Computation* 3, no. 1 (2011): 5-24.

[33] Laurent Itti, "Quantifying the Contribution of Low-Level Saliency to Human Eye Movements in Dynamic Scenes," *Visual Cognition* 12, no. 6 (2005): 1093-1123.

[34] Smith, "Watching You."

[35] For review of empirical studies of film cognition see Tim J. Smith, Daniel Levin, and James E. Cutting, "A Window on Reality: Perceiving Edited Moving Images," *Current Directions in Psychological Science* 21, no. 2 (2012): 107-113.

[36] Harry McGurk and John MacDonald, "Hearing Lips and Seeing Voices," *Nature* 264, no. 5588 (1976): 746-748.

[37] Michel Chion, *Audio-Vision: Sound on Screen*, ed. and trans. Claudia Gorbman (New York: Columbia University Press, 1990).

[38] M. H. Giard and F. Peronnet, "Auditory-Visual Integration During Multimodal Object Recognition in Humans: A Behavioral and Electrophysical Study," *Journal of Cognitive Neuroscience* 11, no. 5 (1999): 473-490.

[39] David Bordwell and Kristin Thompson, *Film Art: An Introduction*, 6th ed. (New York: McGraw Hill, 2001), 333.

[40] Cliodhna Quigley, Selim Onat, Sue Harding, Martin Cooke, and Peter König, "Audio-Visual Integration During Overt Visual Attention," *Journal of Eye Movement Research* 1, no. 2 (2008): 1-17.

[41] Melissa L.-H. Võ, Tim J. Smith, Parag K. Mital, and John M. Henderson, "Do the Eyes Really Have it? Dynamic Allocation of Attention When Viewing Moving Faces," *Journal of Vision* 12, no. 13 (2012): 1-14.

[42] Eric Vatikiotis-Bateson, Inge-Marie Eigsti, Sumio Yano, and Kevin G. Munhall, "Eye Movement of Perceivers During Audiovisual Speech Perception," *Perception and Psychophysics* 60, no. 6 (1998): 926-940.

[43] Charles Spence and Jon Driver, "Audiovisual Links in Exogenous Covert Spatial Orienting," *Perception and Psychophysics* 59, no. 1 (1997): 1-22; John J. McDonald, Wolfgang A. Teder-Sälejärvi, and Steven A. Hillyard, "Involuntary Orienting to Sound Improves Visual Perception," *Nature* 407 (2000): 906-908.

[44] Erik Van der Burg, Christian N. L. Olivers, Adelbert W. Bronkhorst, and Jan Theeuwes, "Pip and Pop: Nonspatial Auditory Signals Improve Spatial Visual Search," *Journal of Experimental Psychology: Human Perception and Performance* 34, no. 5 (2008): 1053-1065.

[45] Antoine Coutrot, Nathalie Guyader, Gelu Ionescu, and Alice Caplier, " Influence of Soundtrack on Eye Movements During Video Exploration*," Journal of Eye Movement Research* 5, no. 4.2 (2012): 1-10.

[46] Võ, et al.

[47] Anna Vilaró, Andrew T. Duchowski, Pilar Orero, Tom Grindinger, Stephen Tetreault, and Elena di Giovanni, " How Sound is the Pear Tree Story? Testing the Effect of Varying Audio Stimuli on Visual Attention Distribution," *Perspectives: Studies in Translatology* 20, no. 1 (2012): 55-65.

[48] J. Vroomen and B. de Gelder, "Sound Enhances Visual Perception: Cross-Modal Effects of Auditory Organization on Vision," *Journal of Experimental Psychology: Human Perception and Performance* 26, no. 5 (2000): 1583-1590.

[49] Jared E. Miller, Laura A. Carlson, and J. Devin McAuley, "When What You Hear Influences When You See: Listening to an Auditory Rhythm Influences the Temporal Allocation of Visual Attention," *Psychological Science* 24, no. 1 (2013): 11-18.

[50] Vroomen and de Gelder.

[51] Eisenstein, *Film Sense*, 137.

[52] Twenty-six participants were presented the "Battle on Ice" sequence (Chapter 10, 52:06-55:17; ripped from DVD as XVID format). Participants were instructed to simply "watch the film clip" whilst their eye movements were recorded using an Eyelink 1000 desk-mounted eyetracker (SR Research). The film was presented on a 21 inch CRT monitor (720x576 pixel resolution DVD quality; viewing angle = 36.44 degrees) at a distance of 60cm with participant head stabilised on a chinrest. Audio was presented on Sennheiser stereo headphones converted from

original mono audio recording (i.e. left/right audio channels were identical). Participants were randomly allocated to one of two groups: with original audio (13 participants) and without audio (13 participants). Gaze data was analysed using Data Viewer (SR Research) to generate peak-through heat-maps of fixation distributions for each shot (
and 3) and parsed into fixations and saccades using standard filters (Dave Stampe, "Heuristic Filtering and Reliable Calibration Methods for Video-Based Pupil-Tracking Systems," *Behaviour Research Methods, Instruments, and Computers* 25, no. 2 (1993): 137-142). Note we do not have to be concerned about smooth pursuit eye movements in this sequence as objects are mostly stationary. Raw gaze data was also parsed into frame-based coordinates for each participant and mapped back on to the original sequence using CARPE (Mital et al.).

[53] Coutrot et al.

[54] DVClab account on Youtube. Video entitled "Audiovisual correspondences in Alexander Nevsky": http://www.youtube.com/watch?v=KBKRSFP9KUM. The center of each viewer's gaze is represented by a small green eye and surrounded by a small Gaussian blob (4 degrees in diameter; roughly the size of the foveal region). The more concentrated the gaze is on one part of the image the hotter the gaussians appear (hence this is referred to as a gaze "heat-map"). This visualisation allows us to gain a qualitative insight into the distribution of gaze throughout the sequence and identify centers of interest (that is, where gaze is most clustered).

[55] Michael Dorr, Thomas Martinetz, Karl R. Gegenfurtner, and Erhardt Barth, "Variability of Eye Movements When Viewing Dynamic Natural Scenes," *Journal of Vision* 10, no. 10.28 (2010): 1-17; Mital et al.

[56] The center of the face; Võ et al.

[57] Eisenstein, *Film Sense*, 138.

[58] Mital et al.

[59] Carmi and Itti.

[60] Wolfe and Horowitz.

[61] Richard D. Pepperman, *The Eye is Quicker: Film Editing: Making a Good Film Better* (Studio City CA: Michael Wiese Productions, 2004), 11.

[62] Bordwell and Thompson.

[63] Võ et al.; Coutrot et al.; Vilaró et al.

[64] Chion; Murch, *In The Blink Of An Eye.*